Trimmed Likelihood-based Estimation in Binary Regression Models

Cizek, P.

Publication date:
2005

CentER

Discussion Paper

No. 2005–108

TRIMMED LIKELIHOOD-BASED ESTIMATION IN BINARY REGRESSION MODELS

By Pavel Čížek

September 2005

ISSN 0924-7815

Tilburg University

Trimmed likelihood-based estimation in binary regression models

Pavel Čížek

Abstract: The binary-choice regression models such as probit and logit are typically estimated by the maximum likelihood method. To improve its robustness, various M-estimation based procedures were proposed, which however require bias corrections to achieve consistency and their resistance to outliers is relatively low. On the contrary, traditional high-breakdown point methods such as maximum trimmed likelihood are not applicable since they induce the separation of data and thus non-identification of estimates by trimming observations. We propose a new robust estimator of binary-choice models based on a maximum symmetrically trimmed likelihood estimator. It is proved to be identified and consistent, and additionally, it does not create separation in the space of explanatory variables as the existing maximum trimmed likelihood. We also discuss asymptotic and robust properties of the proposed method and compare all methods by means of Monte Carlo simulations.

Zusammenfassung: Regressionsmodelle mit diskreten abhängigen Variablen, z.B. probit und logit, werden typischerweise durch das Maximum-Likelihood Prinzip geschätzt. Wegen ihrer niedrigeren Robustheit wurden verschiedene M-Schätzer von binären Modellen vorgeschlagen, die aber auch asymptotisch auf Verzerrung korigiert werden müssen und nicht besonders robust sind. Andererseits sind hoch-robuste Methoden der linearen Regression, z.B. die Maximum-Trimmed-Likelihood Methode, nicht anwendbar, weil sie nicht identifiziert werden können. Hier konstruieren wir einen robusten Schätzer für binäre Regressionsmodelle, der auf eine symmetrisch beschneidete Maximum-Likelihood Methode basiert. Der Schätzer ist bewiesen, identifiziert und konzistent zu sein. Wir diskutieren auch seine robusten Eigenschaften und vergleichen ihm mit anderen bekannten robusten Methoden durch die Monte Carlo Simulationen.

Keywords: binary-choice regression, maximum likelihood, robust estimation, trimming.
JEL codes: C13, C25

1 Introduction

The binary-choice regression models such as probit and logit are used to describe the effect of explanatory variables $x_i \in R^p$ on a binary response variable $y_i \in \{0, 1\}, i = 1, \ldots, n$:

$$P(y_i = 1|x_i) = F(x_i^\top \beta), \tag{1}$$

where F is a known link function (e.g., the standard normal distribution function Φ for probit and the logistic distribution function Λ for logit) and $\beta \in R^p$ is a vector of

unknown parameters. Applications include estimating probability of a firm's bankruptcy and modeling decisions to work, retire, or have children.

Model (1) is typically estimated by the maximum likelihood estimator (MLE), which is defined by

$$\hat{\beta}^{MLE} = \arg\max_{\beta} \sum_{i=1}^{n} l(y_i, x_i; \beta), \tag{2}$$

where the likelihood contributions are

$$l(y_i, x_i; \beta) = y_i \ln F(x_i^\top \beta) + (1 - y_i) \ln\{1 - F(x_i^\top \beta)\}. \tag{3}$$

This estimator is identified only if the two parts of data given by the values of the response variable, $\{x_i|y_i = 1\}$ and $\{x_i|y_i = 0\}$, are not separated in the space of explanatory variables (Albert and Anderson, 1984). MLE is also asymptotically normal and efficient, but it can behave rather poorly if data are contaminated (Croux et al., 2002); for example, if data contain misclassified observations with extreme values of explanatory variables or exhibit an uknown form of heteroscedasticity. Several robust alternatives have been therefore proposed and studied.

In this context, traditional robust (high-breakdown point) methods such as nonlinear least trimmed squares (LTS; Stromberg and Ruppert, 1992; Čížek, 2005) and maximum trimmed likelihood (MTLE; Müller and Neykov, 2003) are not generally applicable since, by trimming observations, they induce the separation of data and thus non-identification of estimates. The only exception are data sets containing large strata, where the number of observations at any observed point x_i grows with sample size (Christman, 1994). Therefore, most recent results rely on M-estimation to achieve robustness: the likelihood contribution function $l(y, x; \beta)$ is replaced by another function $\phi(y, x; \beta)$, which is bounded and possibly contains "weighting" part $w(x; \gamma)$ depending only on the explanatory variables x and some nuisance parameters γ. Recent examples include Copas (1988), Carroll and Pederson (1993), Bianco and Yohai (1996), Kordzakhia et al. (2001), Croux and Haesbroeck (2003), and Gervini (2005).

The described approach based on M-estimation has in many cases two important deficiencies: asymptotic bias causing inconsistency and relatively low robustness. First, the inconsistency of these estimators was noted, for example, by Carroll and Pederson (1993) and can be remedied only by finding and including a bias-correction term into the objective function of a respective estimator (see Bianco and Yohai, 1996, for instance). A disadvantage stemming from this approach lies in low flexibility of such procedures: consistent robust estimators are often designed for logit and their adaptation to other (more flexible) models like in Hausmann et al. (1998) can require redesign of the estimation procedure. Next, the low robustness of M-estimators to misclassified observations with extreme value of explanatory variables was observed and remedied, for example, by Croux and Haesbroeck (2003) and Gervini (2005). A typical remedy unfortunately relies on simple downweighting of distant observations in the space of explanatory variables irrespective to whether they are misspecified or not and to what influence they have on the model.

We propose a new robust estimator of binary-choice models. Even though it relies on a symmetrically trimmed form of maximum likelihood estimator, it is proved to be

identified and consistent in a very general setting. Thus, it does not exhibit any asymptotical bias, it is widely applicable, and additionally, it does not create separation in the space of explanatory variables as LTS and MTLE do. In the rest of this paper, we first identify the source of non-identification of MTLE caused by trimming and motivate a solution in Section 2. Further, we discuss conditions under which the proposed solution is consistent in Section 3 and we mention some important robust properties in Section 4. Finally, we compare the proposed method with some existing solutions using Monte Carlo simulations in Section 5.

2 Identification

Let us first demonstrate why the classical trimmed estimators such as MTLE are not correctly identified in model (1), which will later motivate our proposal. Maximum trimmed likelihood estimator (MTLE) is defined by

$$\hat{\beta}^{(MTLE,h_n)} = \arg\max_{\beta \in B} \sum_{j=1}^{n} \ln l(x_i, y_i; \beta) \cdot I\left(\ln l(x_i, y_i; \beta) \geq \ln l_{[n-h_n+1]}(x_i, y_i; \beta)\right), \quad (4)$$

where $l_{[j]}(x_i, y_i; \beta)$ denotes the jth order statistics of likelihood contributions $l(x_i, y_i; \beta)$, $i = 1, \ldots, n$, and $h_n \in \{[n/2] + 1, \ldots, n\}$ is the trimming constant. The trimming constant determines how many observations h_n are kept in the objective function and how many observations $n - h_n$ are excluded from estimation to protect the estimator against errors in data. The rule used for trimming in (4) is described by the indicator function

$$I\left(\ln l(x_i, y_i; \beta) \geq \ln l_{[n-h_n+1]}(x_i, y_i; \beta)\right)$$

and keeps in the objective function the h_n "most likely" observations, that is, h_n observations with the largest likelihood.

If MTLE as an extremum estimator is identified, the expectation of its objective function (see Čížek, 2004, for derivation)

$$IC(\beta) = E\left[\ln l(x_i, y_i; \beta) \cdot I\left(\ln l(x_i, y_i; \beta) \geq q_\lambda(\beta)\right)\right]$$

has to have a maximum at the true value β_0 of parameter vector β; $q_\lambda(\beta)$ refers here to the λ-quantile of the distribution of $\ln l(x_i, y_i; \beta)$, where $\lambda = 1 - \lim_{n\to\infty} h_n/n$. Therefore, if the MTLE estimator is identified, the first-order conditions $\partial IC(\beta)/\partial\beta = 0$ should hold at β_0.

To verify the first-order conditions, let f denote the density function corresponding to F in (1) and note that (3) and the law of iterated expectation implies (see Čížek, 2004, for details)

$$\begin{aligned}\frac{\partial IC(\beta_0)}{\partial\beta} =& E\left[\left\{\frac{y_i f(x_i^\top\beta_0)}{F(x_i^\top\beta_0)}x_i - \frac{(1-y_i)f(x_i^\top\beta_0)}{1-F(x_i^\top\beta_0)}x_i\right\} I\left(\ln l(x_i, y_i; \beta_0) \geq q_\lambda(\beta_0)\right)\right] \\ =& E_x\left[P(y_i = 1|x_i)\frac{f(x_i^\top\beta_0)}{F(x_i^\top\beta_0)}x_i I\left(\ln l(x_i, 1; \beta_0) \geq q_\lambda(\beta_0)\right)\right] \end{aligned} \quad (5)$$

$$
\begin{aligned}
&-E_x\left[P(y_i=0|x_i)\frac{f(x_i^\top\beta_0)}{1-F(x_i^\top\beta_0)}x_i I\left(\ln l(x_i,0;\beta_0)\geq q_\lambda(\beta_0)\right)\right] \qquad (6)\\
&=E_x\left\{f(x_i^\top\beta_0)x_i\Big[I\left(\ln F(x_i^\top\beta_0)\geq q_\lambda(\beta_0)\right)-I\left(\ln\{1-F(x_i^\top\beta_0)\}\geq q_\lambda(\beta_0)\right)\Big]\right\}.
\end{aligned}
$$

Hence, the first-order condition is satisfied in general only if it holds for all possible values of the random vector x that

$$
I\left(\ln F(x^\top\beta_0)\geq q_\lambda(\beta_0)\right)=I\left(\ln\{1-F(x^\top\beta_0)\}\geq q_\lambda(\beta_0)\right); \qquad (7)
$$

that is, only in the case of the MLE objective function with no trimming, $q_\lambda(\beta_0)=-\infty$ and $\lambda=0$, and in the case of the constantly zero objective function, $q_\lambda(\beta_0)=0$ and $\lambda=1$. Thus, the MTLE estimator is not identified at any $\lambda\in(0,1)$.

On the other hand, this derivation hints that the necessary identification condition would hold if the rule used for trimming observations has the same form both in (5) and (6). In other words, the first-order condition would hold if the trimming rule is independent of the value y_i, which motivates the following proposal: instead of the log-likelihood contributions, let us compare the minimum of the log-likelihood contributions taken over all possible values of $y_i\in\{0,1\}$ and trim observations with low values of

$$
\min\left\{\ln F(x^\top\beta_0),\ln[1-F(x^\top\beta_0)]\right\}.
$$

The resulting *maximum symmetrically trimmed likelihood estimator* (MSTLE) is then defined by

$$
\hat\beta^{(MSTLE,h_n)}=\arg\max_{\beta\in B}\sum_{j=1}^{n}\ln l(x_i,y_i;\beta)\cdot I\left(r(x_i,y_i;\beta)\geq r_{[n-h_n+1]}(x_i,y_i;\beta)\right), \qquad (8)
$$

where $r(x_i,y_i;\beta)=\min\left\{\ln F(x^\top\beta_0),\ln[1-F(x^\top\beta_0)]\right\}$. The first-order conditions for the local identification of the parameter estimates in model (1) are then satisfied as follows from (5)–(6), where $q_\lambda(\beta)$ has to refer now to the λ-quantile of the distribution G_β of $r(x_i,y_i;\beta)$. Complete verification of both the first-order and second-order identification conditions is done in Čížek (2001).

3 Asymptotic properties

The maximum symmetrically trimmed likelihood estimator defined by (8) can be identified in binary-choice models as argued in Section 2. In this section, we demonstrate that it is also consistent under rather general conditions, and therefore, does not require any asymptotic bias correction as many existing M-estimators. We first discuss the sufficient conditions for the consistency of MSTLE and provide the corresponding theoretical result. Later, we mention additional conditions that might be necessary to prove $\sqrt{n}$-consistency and asymptotic normality of this estimator.

The assumptions sufficient for the MSTLE consistency form three groups: distributional assumptions D, assumptions F concerning the MSTLE objective function, and identification assumptions I.

D Let random variables $\{y_i, x_i\}_{i \in N}$ form an independent and identically distributed sequence of random vectors with finite second moments. Further, assume that the distribution function G_β of $r(x_i, y_i; \beta)$ is absolutely continuous with density g_β for any $\beta \in B$ and that it holds for $m_G = \inf_{\beta \in B} q_\lambda(\beta)$ and $M_G = \sup_{\beta \in B} q_\lambda(\beta)$ that

$$M_{gg} = \sup_{\beta \in B} \sup_{z \in (m_G - \delta, M_G + \delta)} g_\beta(z) < \infty \tag{9}$$

and

$$m_{gg} = \inf_{\beta \in B} \inf_{z \in (-\delta, \delta)} g_\beta \left(q_\lambda(\beta) + z \right) > 0 \tag{10}$$

for some $\delta > 0$.

F Let $l(x_i, y_i; \beta)$ be continuous (uniformly over any compact subset of the support of (x_i, y_i)) in $\beta \in B$. Further, let expectation $E \sup_{\beta \in B} |l(x_i, y_i; \beta)|^{1+\delta}$ exist and be finite for some $\delta > 0$.

I Let B be a compact parametric space, and for any $\varepsilon > 0$ and $U(\beta_0, \varepsilon)$ such that $B \backslash U(\beta_0, \varepsilon)$ is compact, let $\alpha(\varepsilon) > 0$ exist such that it holds

$$\begin{aligned} \min_{\beta \in B \backslash U(\beta_0, \varepsilon)} & E \left[l(x_i, y_i; \beta) \cdot I \left(r(x_i, y_i; \beta) \leq q_\lambda(\beta) \right) \right] \\ & - E \left[l(x_i, y_i; \beta_0) \cdot I \left(r(x_i, y_i; \beta_0) \leq q_\lambda(\beta_0) \right) \right] > \alpha(\varepsilon). \end{aligned}$$

Whereas some assumptions are well-known from the literature, such as the existence of the finite first or second moments of random variables and the identification assumptions I mentioned already in Section 2, there is one less usual regularity assumption. It stems from the generality of the model specification, which does not require anything but continuity of the link function F. Assumptions (9) and (10) formalize two things: (i) the density function g_β has to be bounded uniformly in $\beta \in B$, which prevents distribution G_β to be arbitrarily close to a discrete one within the parametric space B; and (ii) the density function g_β has to be positive in a neighborhood of the λ-quantile of G_β, that is, around the chosen "trimming" point of the $r(x_i, y_i; \beta)$ distribution. This type of assumptions is standard in literature on asymptotics of trimmed estimators, see Čížek (2005) for more details.

Under these conditions, it is possible to prove the following result.

Theorem 1 *Let Assumptions D, F, and I hold. Then the MSTLE estimator $\hat{\beta}^{(MSTLE, h_n)}$ is weakly consistent, that is, $\hat{\beta}^{(MSTLE, h_n)} \to \beta_0$ in probability as $n \to +\infty$.*

Proof: The theorem is a direct consequence of Čížek (2004, Theorem 2). Q.E.D.

As shown in Čížek (2004), this result can be extended to derive the $\sqrt{n}$-rate of convergence of the MSTLE estimator if additional assumptions regarding differentiability of $l(x_i, y_i; \beta)$ and some other regularity assumptions are satisfied. Even though it is seems that the same conditions should be sufficient for proving the asymptotic normality of MSTLE, no such result is currently available.

4 Robust properties

After proving that MSTLE is a valid estimator of model (1), we concetrate now on the robustness of the proposed solution. Traditionally, the global robustness of an estimator is measured by the breakdown point. It can be defined as the largest fraction $(m-1)/n$ of sample observations that can be arbitrarily changed without making the estimator "useless" (and naturally, changing then m observations in a right way can make the estimator "useless"), that is, without making estimator a constant, non-random function (Genton and Lucas, 2003).

One of the first results concerning the breakdown point in the binary-choice regression is by Christmann (1994), who shows that the breakdown point ε_n^* of most estimators is design (sample) specific,

$$\varepsilon_n^* \leq \frac{1}{n}\left[\min\left\{\sum_{i=1}^{n} y_i, n - \sum_{i=1}^{n} y_i\right\} - 1\right],$$

and depends on the relative number of observations with responses $y_i = 1$ and $y_i = 0$, respectively. The following theorem complements this general result by providing upper bounds for the breakdown point of MSTLE. They are not sample specific and indicate that, contrary to linear regression, trimming more observations does not necessarily result in a higher breakdown point.

Theorem 2 *The breakdown point of MSTLE estimator (8) with trimming* $h_n \in \{[n/2]+1, \ldots, n\}$ *is in model (1) bounded by* $\varepsilon_n^* \leq [h_n/2]/n$.

Proof: Consider a sample $(x_i, y_i)_{i=1}^n$ and define a contaminated sample $(x_i^*, y_i^*) = (x_i, y_i)$ for $i = 1, \ldots, n - [h_n/2] - 1$ and $(x_{n-i}^*, y_{n-i}^*) = (x_i, 1 - y_i)$ for $i = 1, \ldots, [h_n/2] + 1$. Thus, we changed only $[h_n/2]+1$ observations so that the new sample contains $[h_n/2]+1$ pairs of observation with identical values x_i and complementary values y_i. The MSTLE estimator applied to (x_i^*, y_i^*) trims all non-paired observations and results in $\hat{\beta} = 0$ because both the joint likelihood and trimming rule $r(x_i^*, y_i^*; \beta) = \ln(1/2)$ of all paired observations reach its maximum at $\beta = 0$. Thus, all other (non-paired) observations are trimmed from the objective function. Q.E.D.

On the one hand, the breakdown point is thus bounded by $(n-h_n)/n$ because $n-h_n$ determines the number of observations that can be trimmed from the objective function. On the other hand, misspecification of the values of the dependent variable described in Theorem 2 imposes another bound $[h_n/2]/n$. Consequently, trimming constant h_n should not be chosen smaller than $h_n = [(2n)/3]$, which follows from equating the two bounds, $(n - h_n)/n = h_n/(2n)$, and indicates $\varepsilon_n^* \leq 1/3$. Due to further data-specific limits on the breakdown point (as in Christmann, 1994), $h_n \geq [(3n)/4]$ will probably be a realistic choice in applications.

Finally, note the breakdown point describes a method's behavior only in the extreme situation of its failure. The influence of a point-mass contamination at various locations on the estimation can be however quantified by the so-called bias curve. Because it is difficult to obtain an analytic expression for the bias curve, we will evaluate it by means of Monte Carlo simulations in Section 5 and compare it with bias curves of other existing estimators.

5 Simulation study

To compare the performance of various methods for estimating binary-choice regression models in finite samples, Monte Carlo simulations are used. In this section, we compare the proposed MSTLE method with MLE and the Bianco and Yohai (1996) estimator (BYE), which is based on a bias-corrected M-estimator and was implemented by Croux and Haesbroeck (2003). We also consider weighted forms of MLE and BYE based on weights defined by

W $w_i = I(RD_i^2 \leq \chi^2_{p,0.975})$, where $\chi^2_{p,0.975}$ denotes the 97.5% quantile of χ^2 distribution with p degrees of freedom and RD_i represents the Mahalanobis distance of the ith observation based on the robust MCD estimate of location and covariance (see Croux and Haesbroeck, 2003, for details);

WT $w_i = \min\{c, \exp(r(x_i, y_i; \beta))\} = \min\left\{c, F(x^\top \beta_0), 1 - F(x^\top \beta_0)\right\}$, where $r(x_i, y_i; \beta)$ is the rule used for trimming in (8) and $c = 0.1$, for instance.

The first choice defines weights just by the position of observations in the space of explanatory variables and downweights all distant observations. It is frequently used in the literature (e.g., Croux and Haesbroeck, 2003; Gervini, 2005). The latter choice relies on the initial robust fit by MSTLE and downweights only observations with low values of trimming rule $r(x_i, y_i; \beta)$. The precise choice of weights is arbitrary at this moment and optimal weighting scheme has to be further researched.

As BYE is currently implemented only for logit, we compare all methods using a logistic model as a data-generating process. Specifically, we generate two explanatory variables $x_1, x_2 \sim N(0, 1)$, and for a given parameter vector $b = (b_0, b_1, b_2)$, we define $y = I(b_0 + b_1 x_1 + b_2 x_2 + \varepsilon \geq 0)$, where $\varepsilon \sim \Lambda(0, 1)$ ($N(\mu, \sigma)$ and $\Lambda(\mu, s)$ refer to the Gaussian and logistic distributions, respectively). If a generated data set is not further modified, we refer to it as CLEAN. Next, to examine robust properties of all estimators, we also use contaminated data: a given fraction $\alpha \in (0, 1)$ of observations is shifted by $(\Delta_1, \Delta_2) \in R^2$ and misclassified, which corresponds to transformations $x_1^* = x_1 + \Delta_1, x_2^* = x_2 + \Delta_2$, and $y^* = I(b_0 + b_1 x_1^* + b_2 x_2^* < 0)$. Such data sets are referred to as OUTLIERS(α; Δ_1, Δ_2). Finally, to estimate bias curves of all estimators, we use data with a point-mass contamination: a given fraction $\alpha \in (0, 1)$ of observations is set to (Δ_1, Δ_2) and misclassified, which corresponds to setting $x_1^* = \Delta_1, x_2^* = \Delta_2$, and $y^* = I(b_0 + b_1 x_1^* + b_2 x_2^* < 0)$. These data sets are denoted POINTCONT(α; Δ_1, Δ_2).

Let us note that the results discussed in this section are obtained for sample sizes $n = 100$ observations, trimming constant $h_n = 75$, and 500 simulations. Although we also experimented with larger sample sizes, it seems that the performance of MSTLE at smaller samples is worse relative to other methods than at larger samples, and therefore, we present less favorable results for MSTLE.

5.1 Bias curve

To quantify influence of data contamination on estimation, we evaluate the bias curves of all discussed estimators in the logistic model with parameters $b = (0.5, 1.0, 0.0)$ with

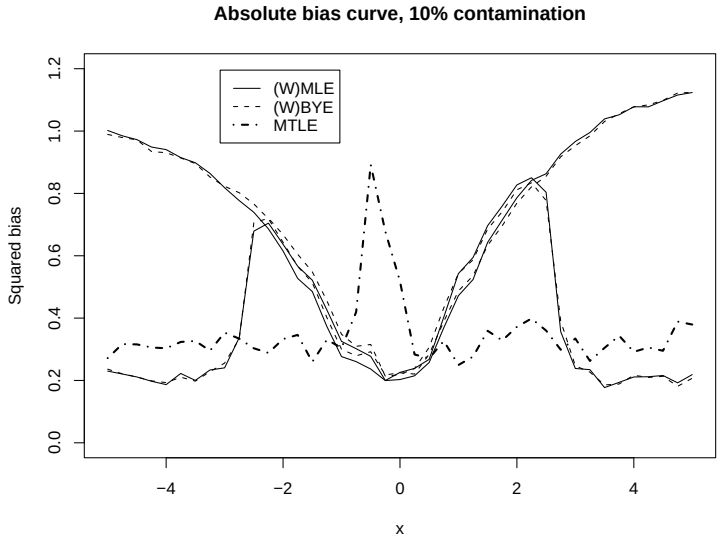

Figure 1: Bias curves of the (W)MLE (solid curves), (W)BYE (dashed curves), and MSTLE (dot-dashed curve) estimators.

10% point-mass contamination at points from interval $(-5, 5)$. This amount to simulating and estimating data POINTCONT$(0.10; x, 0)$ for $x \in (-5, 5)$, which is done here using an equidistant grid with step 0.25. Note that contamination around $x = -0.50$ causes only misclassification, not real outliers.

The results are summarized on Figure 1, which depicts the squared bias of each estimator as a function of contamination point x. First, the standard result indicating low robustness of MLE and BYE is demonstrated here by bias steadily increasing with the increasing distance of contamination point x from the origin. The weighted forms of these estimators, WMLE and WBYE, behave similarly to MLE and BYE for $x^2 \leq \chi^2_{1,0.975}$, but are not influenced by the contamination for $x^2 > \chi^2_{1,0.975}$ because the contaminated observations have then weights equal to zero. The bias curve of MSTLE looks rather differently. On the one hand, it exhibits a comparatively large bias for contamination close to the origin because it uses just pre-specified $h_n = 75$ observations and trims the remaining ones, that is, good ones in this case. On the other hand, the bias of MSTLE is rather small and practically constant for all $x \notin (-1, 0)$, that is, when data contain real outliers. Note that whereas MSTLE performs equally well both in samples with moderate and large outliers, WMLE and WBYE perform well only if outliers are far enough from the correct observations.

5.2 Estimation under contamination

The performance of all methods is now analyzed both under clean and contaminated data sets generated from the logistic model with $b = (0.5, 1.0, -1.0)$. Employed data are CLEAN, OUTLIERS$(0.05; 1.5, -1.5)$, and OUTLIERS$(0.05;\ 5.0, -5.0)$ and the con-

Table 1: Bias and mean squared error (MSE) of (W)MLE, (W)BYE, and MSTLE for clean and contaminated data.

Bias (MSE)	CLEAN	OUTLIERS(0.05;1.5,1.5)	OUTLIERS(0.05;5.0,5.0)
MLE	0.099 (0.261)	0.764 (0.688)	1.396 (2.037)
WMLE	0.103 (0.279)	0.792 (0.749)	0.077 (0.273)
BYE	0.109 (0.281)	0.600 (0.489)	0.960 (1.333)
WBYE	0.111 (0.299)	0.626 (0.537)	0.093 (0.304)
MSTL	0.533 (1.011)	0.539 (0.997)	0.565 (1.041)
WTMLE	0.165 (0.350)	0.018 (0.388)	0.134 (0.382)

tamination level is thus 5%. The absolute value of bias and mean squared error (MSE) for each methods is recorded in Table 1.

First, very high sensitivity of MLE and BYE to outliers is again clearly visible, even though BYE is sligthly less affected by contamination. The corresponding weighted versions, WMLE and WBYE, perform rather well in the case of clean data and data with distant outliers, which can be easily detected and downweighted. Both weighted methods however fail to withstand contaminated data if outliers are not too far from the rest of data. On the contrary, the results of the proposed MSTLE method are practically unaffected by contamination, but are very imprecise; the MSE of MSTLE for clean data is almost four times higher than the MSE of MLE. This well-known inefficiency of trimmed estimators can be overcome by using them only as an initial robust estimator for a more efficient method. In our case, we use MSTLE to construct weights for MLE. The resulting WTMLE estimator is rather close to the performance of existing robust methods for clean data, but is not significantly influenced by the moderate and large outliers.

6 Conclusion

The maximum symmetrically trimmed likelihood estimator proposed in this paper is shown to be applicable in general binary-choice models, consistent, and robust to various kinds of contamination. The combination of these properties is not currently matched by any existing robust method. On the other hand, trimming of observations leads to an inevitable loss of efficiency, which can be however remedied to a large extent by using MSTLE as an initial estimator for weighted MLE. The optimal choice of weights stays as a topic for further research. Similarly, the bias curve of MSTLE indicates that a combination with models accounting for data misspecification (Hausmann et al., 1998) could be beneficial and should be further investigated.

Address: Tilburg University, Department of Econometrics & OR, Room B 616;
P.O. Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail: P.Cizek@uvt.nl

References

A. Albert and J. A. Anderson (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71, 1–10.

A. M. Bianco and V. J. Yohai (1996) Robust estimation in the logistic regression model. In H. Rieder (ed.) *Robust statistics, data analysis, and computer intensive methods*, Lecture notes in statistics 109, Springer, New York, 17–34.

R. J. Carroll and S. Pederson (1993) On robustness in the logistic regression model. *Journal of Royal Statistical Society, Ser. B* 55, 693–706.

A. Christmann (1994) Least median of weighted squares in logistic regression with large strata. *Biometrika* 81, 413–417.

P. Čížek (2001) Robust estimation in nonlinear regression and limited dependent variable models. CERGE-EI Discussion Paper 189/2001, Charles University, Prague.

P. Čížek (2004) General trimmed estimation: robust approach to nonlinear and limited dependent variable models. CentER Discussion Paper 130/2004, Tilburg University, Tilburg.

P. Čížek (2005) Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning & Inference*, to appear.

J. B. Copas (1988) Binary regression models for contamination data. *Journal of Royal Statistical Society, Ser. B* 50, 225–265.

C. Croux, C. Flandre, and G. Haesbroeck (2002) The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters* 60, 377–386.

C. Croux and G. Haesbroeck (2003) Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis* 44, 273–295.

M. G. Genton and A. Lucas (2003) Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of Royal Statistical Society, Ser. B* 65, 81–94.

D. Gervini (2005) Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference* 131, 297–311.

J. A. Hausman, J. Abrevaya, and F. M. Scott-Morton (1998) Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87, 239–269.

N. Kordzakhia, G. D. Mishra, and L. Reiersølmoen (2001) Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference* 98, 211–223.

C. H. Müller and N. M. Neykov (2003) Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116, 503–519.

A. J. Stromberg and D. Ruppert (1992) Breakdown in nonlinear regression. *Journal of American Statistical Association* 87, 991–997.